DOCUMENT RESUME

ED 074 758 EM 010 946

AUTHOR Kniefel, David R.
TITLE QUEST: A Conversational Access to Computerized Searches of the ERIC System.
PUB DATE 1 Mar 73
NOTE 18p.; Paper presented at the American Educational Research Association Meeting (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Computer Programs; Flow Charts; *Information Processing; *Information Retrieval; Information Sources; *Information Systems; Relevance (Information Retrieval); *Search Strategies
IDENTIFIERS *Educational Resources Information Center; ERIC; North Carolina Science Technology Research Center; QUEST; STRC-IVS

ABSTRACT

QUEST is an interactive conversational Programming System (CPS) that was developed to serve as a conversational interface between a searcher of the Educational Resources Information Center (ERIC) files and the North Carolina Science and Technology Research Center's Inverted File Search Program (STRC-IVS). This paper describes QUEST: its costs, operational procedures, and problems. The QUEST program has not only proved to be extremely successful cost-wise, but has also become a means of introducing non-computer oriented users to automated information retrieval techniques. The development of QUEST has greatly increased the utilization of the ERIC files. (MC)

ED 074758

NPO
EM

QUEST:

A Conversational Access to Computerized Searches of the ERIC System

David R. Kniefel
North Carolina State University

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

A paper read as a part of the annual meeting of the American Educational Research Association, New Orleans, Louisiana, February 25 - March 1, 1973.

EM 010946

ACKNOWLEDGEMENT

The writer would like to acknowledge and thank the North Carolina Science and Technology Research Center for their cooperation and assistance in this project. Mason Bridenstein and Mary Ann Williamson, STRC staff programmers were responsible for major portions of the QUEST programs. Rodney Wines of NCSU developed the SUBMIT program and aided in the development of QUEST.

INTRODUCTION

The ERIC system is providing educators with a centralized source for the retrieval of educational information. Since its first publication of Research In Education (RIE) in late 1966 and Current Index to Journals In Education (CIJE) in 1969, the ERIC information base has rapidly been expanding. RIE announces the availability (on microfiche) of documents, which prior to ERIC, may have only been available to a very limited audience, e.g., Federal research reports, State Department publications, etc.. CIJE is currently indexing and providing brief annotations of educationally related articles contained in more than 500 journals. The two files, then, provide a broad and up-to-date information base for educators. As of December 31, 1972, RIE and CIJE contained 59,559 and 62,751 citations respectively.

ERIC was developed to capture and catalogue the mass of educational information being produced in a manner that would ease the educator's retrieval of only that information relevant to his needs. The key to the ERIC system is the Thesaurus of ERIC Descriptors. The Thesaurus is essentially a dictionary of synonyms, and serves as a guide to the selection of a term or related terms which have been authorized as a descriptors for indexing documents related to specific concepts. Each citation in the ERIC system is assigned a number of these terms which describe its contents. Even with monthly publications and commulative indexes organized by descriptor subject headings, hand retrieval of information, especially specific information involving a number of interrelated concepts, becomes an awesome task.

The high speed computer provides the tool necessary to gain rapid access to these broad information bases. ERIC was designed with computerized retrieval as one of its basic goals and in 1968 began making its files available on magnetic tape in computer readable form.

Current automated searching techniques are primarily off-line ("batch" processing) or on-line (interactive) processing systems. Off-line systems are highly efficient in terms of computer systems operation and cost. They have the distinct disadvantage of removing the searcher from the search process. In most instances it is necessary to engage an interpreter or information analyst who translates the searcher's request into a language and format acceptable to the computer. These interpreted requests are then keypunched and computer efficiency is gained by "batching" (grouping) these jobs and feeding them into the computer. A further limitation of off-line systems is the delay incurred by interposing an interpreter between the searcher and the computer plus the physical distance between the searcher and the computer.

On-line systems place the searcher in a direct "hands-on" relationship with the information files allowing him to manipulate the information base within the limitations of the system. Furthermore, feedback is instantaneous via telecommunications devices. However, these systems are extremely inefficient in terms of computer systems operation and cost.

The purpose of this paper is to describe QUEST, a conversational interface to an off-line batch search system, which retains the high level of computer and cost efficiency but removes the need for an interpreter between the searcher and the computer.

QUEST

In general terms, QUEST is an interactive, computerized interface between the searcher and the North Carolina Science and Technology Research Center's Inverted File Search Program (STRC-IVS).[1] Essentially, QUEST solicits information from a user through a slow-speed terminal (e.g., a teletype) and creates correctly formated IBM card images suitable for batch processing.

STRC-IVS and ERIC

In June of 1969 the North Carolina State University (NCSU), Center for Occupational Education (COE) purchased the ERIC tapes and entered into a cooperative agreement with the North Carolina Science and Technology Research Center (STRC) to make the ERIC files compatible with the STRC-IVS program. The STRC search system was selected because it had been demonstrated to be highly efficient and was, as a result, being used to access the National Aeronautics and Space Administration's (NASA) and other information files. The first inverted file search of the ERIC system was conducted later in the year and again the STRC-IVS program proved to be highly efficient.

The system is currently operating in the Triangle Universities Computer Center (TUCC) computing environment on an IBM OS 370/165 with the Time Saving Option (TSO) and IBM 2314 disk packs. The 165 ERIC searches (RIE and CIJE) conducted by STRC search analysts since January 1, 1973, have yielded an average of 227 hits (documents selected from files via STRC-IVS) at an average computer cost of $10.00 per search. The search programs are executed at priority = 0, th lowest computer charges in the TUCC environment.

The Development of QUEST

In the fall of 1971 computerized searches of the ERIC files were available to School of Education faculty at NCSU through the Center for Occupational Education. Conducting a search required the submission of a written search request and the payment of a $7.00 fee for the development of the search strategy (term selection and a logic formula) and keypunching. Computer costs were covered by departmental or faculty accounts. While the fee was nominal, departmental budget restrictions limited the extent to which the file could be used.

Instructors who wished to familiarize students with the files and computer access were faced with training students not only to keypunch, but to adhere to the strict formating requirements and the sequencing of the appropriate cards within the program. Speaking from experience, this was an almost impossible task. Thus, the need for a more viable interface between the searcher and the computer was evident.

The School of Education and STRC agreed to cooperate in the development of such an interface. A programmer, engaged part-time by the School of Education, and a staff programmer at STRC devoted a percentage of their time to the development of QUEST. The feasibility of an interactive interface was demonstrated during the spring of 1972, developed and debugged during the summer and a pilot implementation of the system was conducted with success during the fall. Approximately 80 graduate students and faculty (with no computer experience) submitted various searches through the system. Concurrent with those activities was the development of user documentation related to: 1) the development of search strategies and 2) the use of QUEST

and the interactive terminal. Currently any student or faculty member with a valid computer account can access QUEST and conduct a search of the ERIC files.

The QUEST System

QUEST is a CPS (Conversational Programming System) program with attendant CPS file maintenance programs (e.g. ERASE) and one OS/360 assembler program SUBMIT. Figure 1 presents a system flowchart of QUEST and an explanation of the components follows:

QUEST - A conversational program which solicits information from the searcher and creates two direct access disk files of correctly formated card images.

JCL file - is a direct access file of JCL (Job Control Language) card images created by QUEST which initiate STRC-IVS.

Data file - is direct access file of correctly formated data (e.g. terms and equations) on which STRC-IVS operates. The 0 record in this file serves as the index for the file and contains the date entered and the savekeys of the searches contained in the file.

TFILE - QUEST writes the terms entered by the searcher into this direct access file and then based on a logic equation supplied by the searcher reads them from TFILE and writes them in the correct order in the Data file.

ERASE - This conversational file maintenance program erases selected (by savekey or date) searches from the Data file. This allows the system operator to purge the file of data related to completed searches.

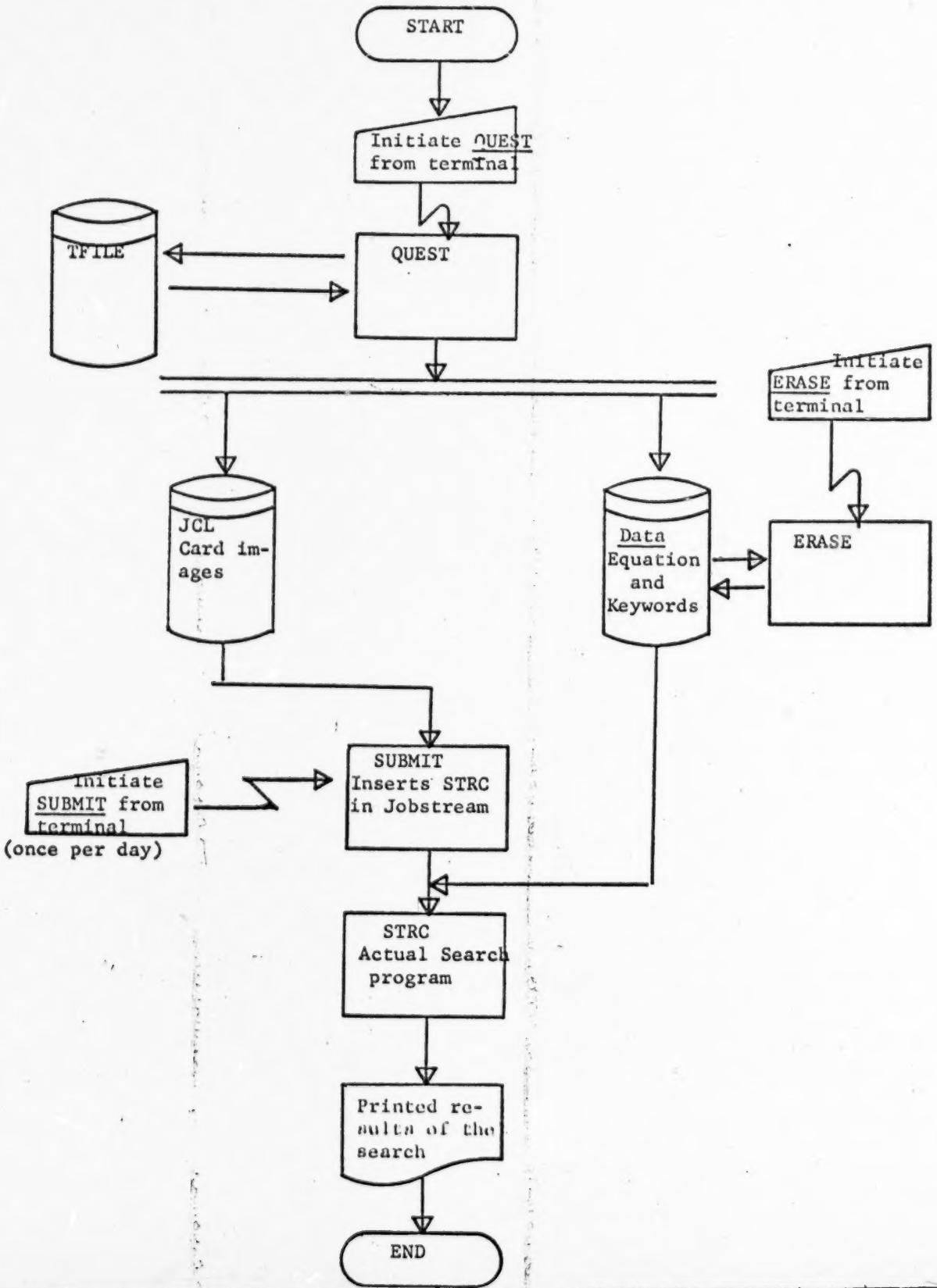
SUBMIT - An OS/360 Assembler Program which reads the JCL file and copies its contents into the jobstream. This program is executed through RJE (Remote Job Entry) from a slow speed terminal once a day.

The output from the batch (STRC-IVS) processing is routed to the high speed printer located in the NCSU, Computer Center. Results of a search entered on one day are available in the School of Education, Computer Facility by 9:30 a.m. on the following day.

SYSTEM-QUEST
Date Prepared-1/31/73
Prepared by- Robert Hamer

FIGURE 1

QUEST SYSTEM FLOWCHART



Computer costs for interactive slow-speed I/O in the TUCC, NCSU environment are currently computed in terms of CPU @ .41¢/second and connect times (port charges, I/O etc.) @ $2.50/hr. Based on a sample of 30 QUEST sessions, current computer charges for the interactive system, with an average of 3.73 CPU seconds per file entry (RIE or CIJE) and 22.45 average minutes of connect time per file entry are $2.43 per search entry per file or $4.86 per RIE and CIJE search. However, batch costs reported by STRC at an average of $10.00 per search of the ERIC system (RIE and CIJE) have been reduced by QUEST users to an average of $4.78 per search with an attendant reduction in the number of documents retrieved (STRC $\overline{X}$ = 227 hits/search, QUEST users $\overline{X}$ = 86.2 hits/search). The reduced printing costs due to the reduced number of hits accounts for much of the apparent discrepancy in batch costs. It is hypothesized that user development of search strategies has led to more specific search equations thereby, reducing the retrieval of irrelevant information more likely when an interpreter is placed between the searcher and the retrieval system. Thus, with an average computer cost of $9.64, QUEST has removed the $7.00 service charge and allowed the direct involvement of the searcher in the retrieval of his information.

The Searcher and Quest

A potential user of the QUEST system is directed to a document entitled Conducting an ERIC Search Using Quest.[2] Part I of this document introduces the searcher to ERIC and leads him through the development of a comprehensive search strategy. This includes familiarizing the user with the Thesaursus of ERIC Descriptors and the process of selecting descriptors related to his information needs. Also, the user is introduced to the STRC RIE & CIJE Dictionaries. These documents present an alphabetical listing of all the terms

used to index citations in RIE and CIJE plus the number of times each term has been assigned to a citation in each file (postings). The postings are used in the construction of the search equation to increase the efficiency of the computer program during the search process. Identifiers (terms not in the Thesaurus that are assigned to specific citations by ERIC information specialists to aid in their retrieval) and terms that were misspelled when they were entered into the ERIC files are also listed for possible inclusion in the search strategy. A brief example of a search strategy follows:

The user is interested in determining the research skills and competencies offered to or required of students engaged in graduate education. A search title is constructed.

TITLE: Research Skills Required in Graduate Education

The user then consults the ERIC Thesaurus and lists the descriptors which are related to his topic. He then finds each term in the STRC dictionaries and lists the postings for each term in each file and adds any pertinent misspelled terms or identifiers.

Search Terms and Postings:	RIE	CIJE
1. Research Skills	76	42
2. Graduate Students	5	25
3. Graduate Study	83	97
4. Research Tools	69	84
5. Resrch Tools	0	1
	233	249

The last step is to construct the logic equation using the numbers assigned to the terms for each file to be searched. Related terms are grouped together by () and terms and groups are connected with appropriate Boolean operators, and (.), or (+) and not (-). In addition, to

increase the efficiency the computer program operation, the terms within groups and the groups are ordered. The term with the lowest number of postings is placed first in a group and the group with the lowest total number of postings comes first in the equation. Also terms with 0 postings in a file are not included in that files equation.

Search Equations:

RIE	(2 + 3) . (4 + 1)
CIJE	(5 + 1 + 4) . (2 + 3)

While the equations appear to be different, the logic of the intersection (and) will result in only citations and abstracts which have one or more of the descriptors in each group assigned to it being returned to the searcher.

Part II, of Conducting an ERIC Search using QUEST familiarizes the user with slow speed terminals i.e. teletype and the IBM 2741, discusses the procedure used to establish communication with the computer (LOGON) and presents a complete script of a terminal session. Figure 2 is a flow chart of the program script indicating the decision points available to the user and the user apparent flow of a terminal session.

A terminal session involving the preceding search strategy for RIE is provided as an example of a typical user interaction with QUEST. Computer initiated text is in uppercase while user responses is in lower case.

(LOGON PROCEDURE)

```
?load (QUEST)
?xeq
ARE YOU ENTERING A NEW SEARCH
?Yes
YOUR SAVE KEY FOR THIS SEARCH IS XXX
ENTER YOUR SEARCH TITLE
?research skills required in graduate education
```

FIGURE 2

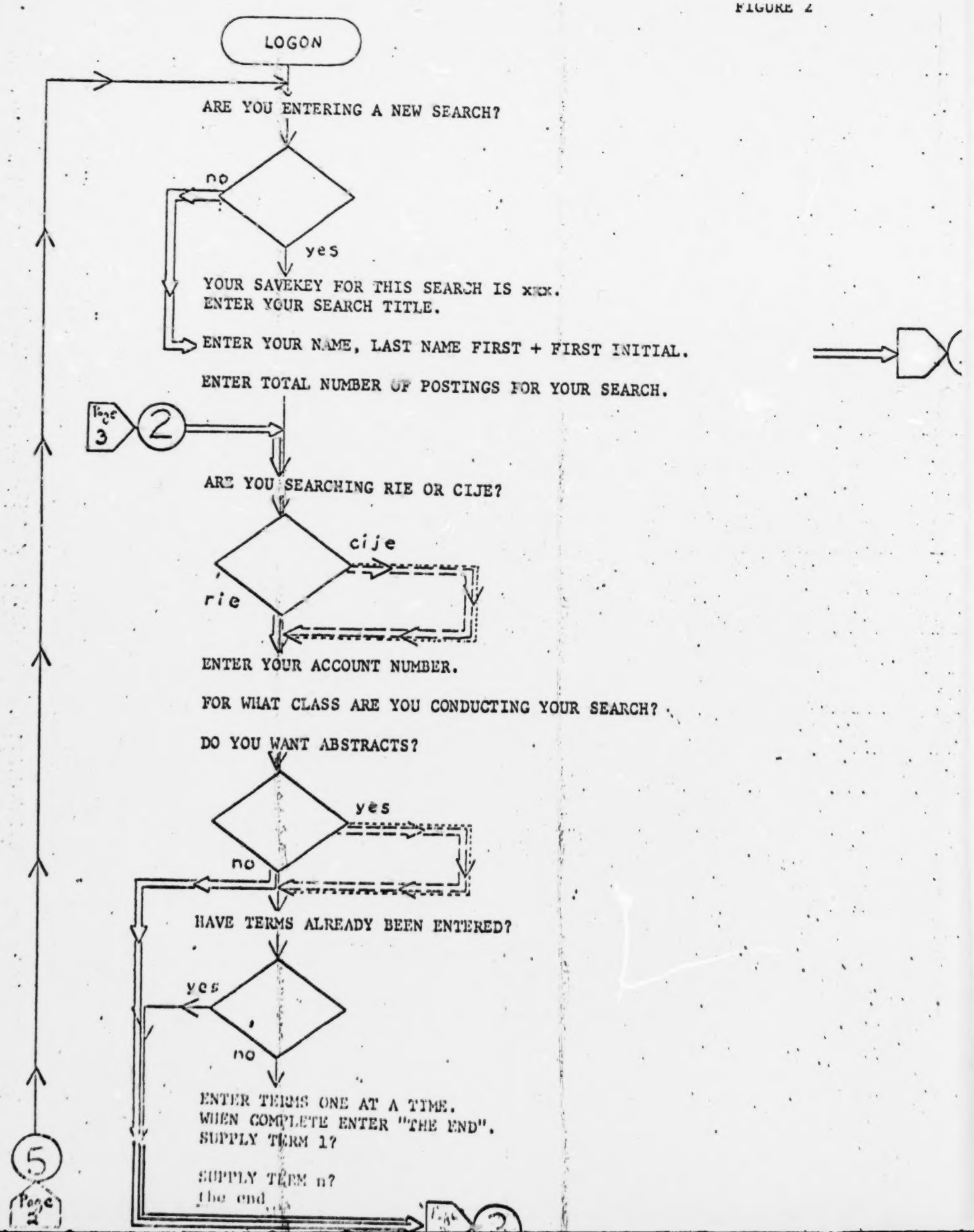

LOGON
ARE YOU ENTERING A NEW SEARCH?
no
yes
YOUR SAVEKEY FOR THIS SEARCH IS xxx.
ENTER YOUR SEARCH TITLE.
ENTER YOUR NAME, LAST NAME FIRST + FIRST INITIAL.
ENTER TOTAL NUMBER OF POSTINGS FOR YOUR SEARCH.
Page 3
2
ARE YOU SEARCHING RIE OR CIJE?
cije
rie
ENTER YOUR ACCOUNT NUMBER.
FOR WHAT CLASS ARE YOU CONDUCTING YOUR SEARCH?
DO YOU WANT ABSTRACTS?
yes
no
HAVE TERMS ALREADY BEEN ENTERED?
yes
no
ENTER TERMS ONE AT A TIME.
WHEN COMPLETE ENTER "THE END".
SUPPLY TERM 1?
SUPPLY TERM n?
the end
5
Page 2

FIGURE 2 (cont)

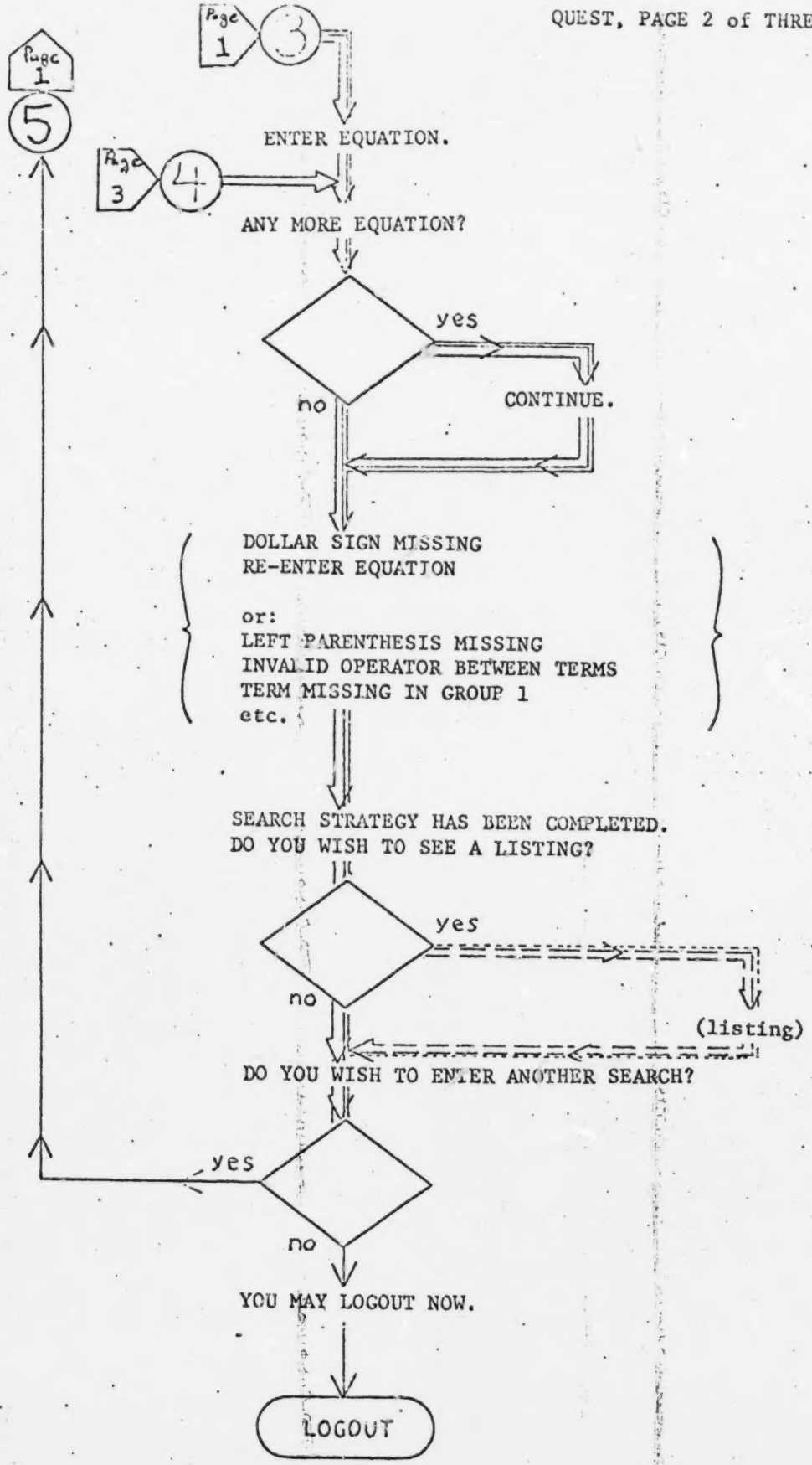
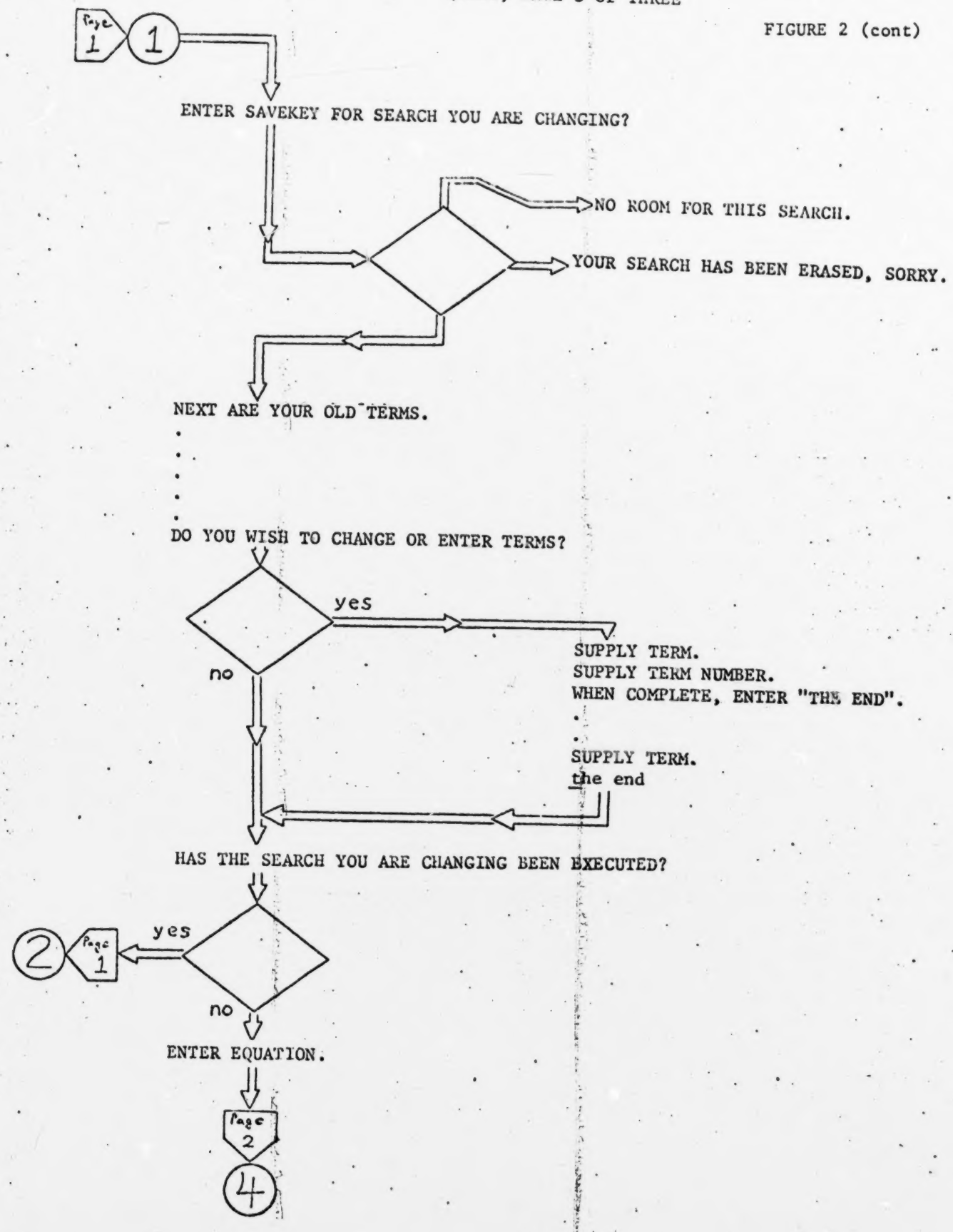


FIGURE 2 (cont)



```
ENTER YOUR NAME, LAST NAME FIRST
?doe John
ENTER THE TOTAL NUMBER OF POSTINGS FOR YOUR SEARCH
?233
ARE YOU SEARCHING RIE OR CIJE
?rie
ENTER YOUR COMPUTER ACCOUNT NUMBER
?ncs.ded.1234
FOR WHAT CLASS ARE YOU CONDUCTING YOUR SEARCH
?ed 615
DO YOU WANT ABSTRACTS
?yes
HAVE TERMS ALREADY BEEN ENTERED
?no
ENTER TERMS ONE AT A TIME.  WHEN COMPLETE ENTER "THE END"
SUPPLY TERM 1
?research skills
SUPPLY TERM 2
?graduate students
SUPPLY TERM 3
?graduate study
SUPPLY TERM 4
?research tools
SUPPLY TERM 5
?resrch tools
SUPPLY TERM 6
?the end
ENTER EQUATION
?(2 + 3) . (4 + 1)
ANY MORE EQUATION
?no
YOUR SEARCH STRATEGY HAS BEEN COMPLETED DO YOU WISH TO SEE A LISTING OF
THIS SEARCH
?no
YOU MAY LOGOUT NOW
(LOGOUT PROCEDURE)
```

The users JCL and search data have been constructed. SUBMIT will be executed and his search will be processed by STRC-IVS. The print-out of the results of his search will be available in the School of Education, Computer Facility by 9:30 a.m. on the day following the users initiation of the QUEST procedure.

CONCLUSIONS

QUEST is a CPS program that was developed to serve as a conversational interface between a searcher of the ERIC information files and STRC-IVS, an inverted file, batch information retrieval system. The approach has proved to be feasible not only from the standpoint of cost but also as a means of introducing non-computer oriented users to automated information retrieval techniques. Furthermore, the procedure removes the need for a human interpreter between the searcher and the computer.

Perhaps the most encouraging aspect of the QUEST development, is the evidence which indicates that the utilization of the ERIC files is increasing dramatically. It is estimated that approximately 150 searches (RIE and CIJE) were conducted by faculty, students and outside organizations during the 1971-72 school year through the C.O.E. Students alone have almost doubled that figure since September of 1972 and all indications are that the utilization will continue to increase as additional faculty and students become familiar with the system. QUEST is not without its limitations. Currently only one user at a time can enter his search strategy. A major problem is one of spelling. The current probability of a searcher receiving output from his first QUEST encounter is about .50 since the computer is highly sensitive to extra spaces and misspelled terms. Another problem resides with the frequent occurrence of computer systems malfunction or downtime, line drops, et.. While these occurrences do not harm QUEST or its files they become highly frustrating to the neophyte user who believes he has broken the machine.

Current activities are directed at rewriting QUEST as a multiple user system. Future plans include: 1) exploration of the feasibility and cost effectiveness of placing the STRC Distionaries on-line for spelling checks and automatic ordering of terms and groups relative to postings; 2) expansion of the availability of the QUEST system through the North Carolina Educational Computing Service and the TUCC telecommunications networks; 3) the development of alternative user documentation strategies, e.g. video tape, film, slide tape, etc..

While on-line, instant retrieval of information is perhaps the most satisfying experience for the user it is extremely expensive in terms of cost and systems operation. It would appear that the QUEST strategy provides the means whereby current, highly efficient, computer techniques can be employed to increase automated retrieval applications for a broad audience of users on a variety of information bases.

FOOTNOTES

[1] Williamson, Mary Ann, The STRC Inverted File Search Program, Technical Report No. 117. North Carolina Science and Technology Research Center, Research Triangle Park, North Carolina 1970.

The STRC search system (STRC-IVS) is written in FORTRAN with one S/360 Assembler subroutine. It provides for computer retrieval of information on subject indexed information files. Performing inverted file I/O, the assembler subroutine uses the Indexed Sequential Access Method (SAM) to provide direct access retrieval capabilities.

Input to the system is a search question in the form of a pseudo Boolean equation with a highly specified format. The equation may contain a variable number of terms arranged within a variable number of groups. Terms and groups are joined by three standard Boolean logical operators; and (.), or (+), and not (-). The groups within an equation are solved (i.e., documents indexed by the terms within a group are identified and saved) and their results are combined according to the Standard Boolean hierarchy of operations (i.e., intersection (and) comes before union (or).

[2] Lowery, Robert and Kniefel, David, Conducting an Eric Search Using Quest. NCSU, School of Education. Spring 1973.